

SCORING PROJECT IN FINANCE

Study of The Purchase of A New Fund of A Financial Institution

Contents

1	Introduction	2
2	Data Description	2
2.1	The target variable - Y	2
2.2	Explanatory variables	3
2.2.1	Descriptive statistics on Binary Variables	3
2.2.2	Descriptive statistics on continuous Variables	4
2.3	Variables with NAs	6
2.4	Categorization of the continuous variables	6
3	Bivariate Analysis	7
3.1	Pearson's chi square χ^2 measure of independance	7
3.2	Proportion of Subscribers by Category	8
4	Prediction	11
4.1	Random Forest	11
4.1.1	Performance on training data	12
4.1.2	Performance on testing data	12
4.2	Support Vector Machine	13
4.2.1	Performance on training data	14
4.2.2	Performance on testing data	14
4.3	Algorithms Comparison	15
5	Conclusion	16

1 Introduction

For the marketing of a new product, any company generally asks itself the question of knowing what is the proportion of customers interested in the product. This is to determine whether the sale of the product will be profitable for the company. In this work, we study the case of a financial institution that wants to market a new Fund to its customers. For this purpose, this institution want to build an economically optimal fund raising campaign. To do so, a test campaign was done on a group of representative agency. Among the 6249 people requested, only 1015 bought the proposed product (about 16%). The aim of this study is to identify the most likely customers to buy this product. Then the company will use it to develop the model Agency traffic building for large amounts and remote subscription for others. The others agencies, will use the previous rule and apply it on their customers.

To reach our objective, that is to estimate the probability to purchase to the new product(Yes/No), we proceed as follows:

- We first gives a description of the variables of the data set together with an analysis of missing values;
- Second, we do a bi-variate analysis between the target variable and others variables; This is to determine the variable that have the most important impact on the target variable;
- Then we use these variables to construct more complex variables;
- Finally, we build two models and conclude.

2 Data Description

We have at our disposal a data set containing 6249 observations and 24 variables.

2.1 The target variable - Y

The target variable is the variable Y on the data set. It is a binary variable that takes the value 1 if a customer purchases the new product and it takes the value 0 otherwise. We recoded it as 1 if the customer purchases the new product and 0 otherwise. The frequency distribution of the target variable Y is described in Table 1 and Figure 1.

Variable	Levels	counts	Proportion	NA
Y	1	1015	16.24%	0%
	9	5234	83.75%	0%

Table 1: Description of the target variable Y

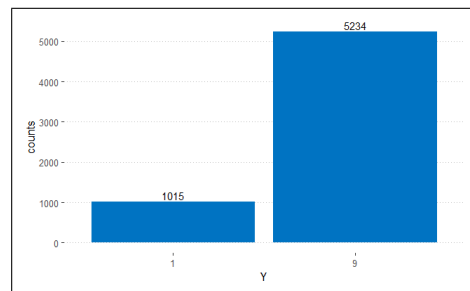


Figure 1: Bar plot of Y

2.2 Explanatory variables

#	Variable	Exact label
1	AGE	Age
2	ANCCDD	Seniority Customer
3	SURFIN	Total saving
4	SOLMOY	Average position of checking account
5	FLUCRE	Credit Flow
6	FLUDEE	Debt flow
7	NBCRE	Volume of crediting transaction
8	NBDEE	Volume of Debiting transaction
9	NBJDE	Volume of day debtor
10	DETVIE	Customer with life insurance Contract (0/1)
11	DETIMMO	Customer with Housing Loan (0/1)
12	DETCONSO	Customer with consumer credit (0/1)
13	DETREV	Customer with credit revolving contract (0/1)
14	SLDLIQ	Savings balance of the customer
15	DETLIQ	Customers with saving account (0/1)
16	CPTTIT	Trading account customer
17	PEA	Customer PEA (0/1)
18	PSOC	Social part
19	SLDTIT	Amount of trading account
20	SLDPEA	Amount of PEA
21	DETBLO	Customers with locked saving account (0/1)
22	SLDBLO	Amount of locked saving
23	NUPER	Ident
24	Y	Y (true responder)

Figure 2: Labels of the explanatory variables

2.2.1 Descriptive statistics on Binary Variables

There are 9 binary variables. Their description is given in Figure 3


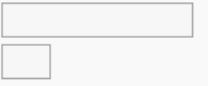
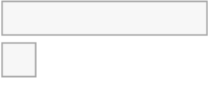
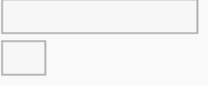

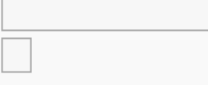
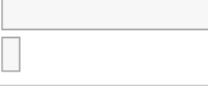

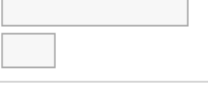
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	DETVIE [factor]	1. 0 2. 1	5356 (85.7%) 893 (14.3%)		6249 (100%)	0 (0%)
2	DETIMMO [factor]	1. 0 2. 1	4972 (79.6%) 1277 (20.4%)		6249 (100%)	0 (0%)
3	DETCOSO [factor]	1. 0 2. 1	5342 (85.5%) 907 (14.5%)		6249 (100%)	0 (0%)
4	DETRV [factor]	1. 0 2. 1	5128 (82.1%) 1121 (17.9%)		6249 (100%)	0 (0%)
5	DETLIQ [factor]	1. 0 2. 1	3468 (55.5%) 2781 (44.5%)		6249 (100%)	0 (0%)
6	CPTTIT [factor]	1. 0 2. 1	5506 (88.1%) 743 (11.9%)		6249 (100%)	0 (0%)
7	PEA [factor]	1. 0 2. 1	5770 (92.3%) 479 (7.7%)		6249 (100%)	0 (0%)
8	PSOC [factor]	1. 0 2. 1	5765 (92.2%) 484 (7.8%)		6249 (100%)	0 (0%)
9	DETBLO [factor]	1. 0 2. 1	4851 (77.6%) 1398 (22.4%)		6249 (100%)	0 (0%)

Figure 3: Descriptive statistics on Binary variables

2.2.2 Descriptive statistics on continuous Variables

There are 14 continuous variables. Their description is given in Figure 4 and 5.

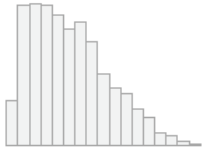
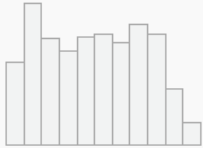


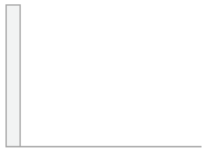

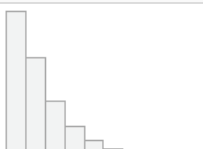
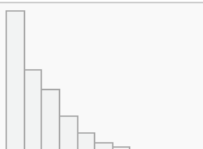

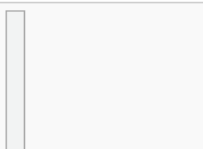
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	AGE [numeric]	Mean (sd) : 43.3 (16.7) min < med < max: 19 < 41 < 100 IQR (CV) : 24 (0.4)	81 distinct values		6249 (100%)	0 (0%)
2	ANCCDD [numeric]	Mean (sd) : 10.4 (5.6) min < med < max: 0 < 10 < 22 IQR (CV) : 10 (0.5)	23 distinct values		6249 (100%)	0 (0%)
3	SURFIN [numeric]	Mean (sd) : 14088.6 (32035.3) min < med < max: 0 < 2626.9 < 654089.3 IQR (CV) : 14602.7 (2.3)	4619 distinct values		6249 (100%)	0 (0%)
4	SOLMOY [numeric]	Mean (sd) : 2009.2 (12495.7) min < med < max: -27295.2 < 768.7 < 632365.9 IQR (CV) : 1780.1 (6.2)	5379 distinct values		6249 (100%)	0 (0%)
5	FLUCRE [numeric]	Mean (sd) : 1784.8 (5648.2) min < med < max: 0 < 1167.7 < 266520.5 IQR (CV) : 1555.7 (3.2)	5500 distinct values		6249 (100%)	0 (0%)
6	FLUDEE [numeric]	Mean (sd) : -1844.3 (3313.3) min < med < max: -89927.2 < -1143.4 < 75.4 IQR (CV) : 1761.5 (-1.8)	5611 distinct values		6249 (100%)	0 (0%)
7	NBCRE [numeric]	Mean (sd) : 3.3 (2.6) min < med < max: 0 < 2.7 < 20 IQR (CV) : 3.3 (0.8)	53 distinct values		6249 (100%)	0 (0%)
8	NBDEE [numeric]	Mean (sd) : 18.2 (16.1) min < med < max: 0 < 14.7 < 110 IQR (CV) : 22.7 (0.9)	240 distinct values		6249 (100%)	0 (0%)
9	NBJDE [numeric]	Mean (sd) : 5.1 (8.6) min < med < max: 0 < 0 < 56.3 IQR (CV) : 7 (1.7)	126 distinct values		6249 (100%)	0 (0%)
10	SLDLIQ [numeric]	Mean (sd) : 4919.2 (7791.8) min < med < max: 0 < 2908.4 < 204349.6 IQR (CV) : 6277.8 (1.6)	5 2692 distinct values		2784 (44.55%)	3465 (55.45%)

Figure 4: Descriptive statistics on the first group of continuous variables




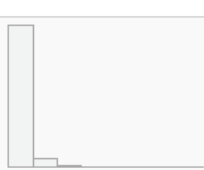
11	SLDTIT [numeric]	Mean (sd) : 18858.9 (31284.4) min < med < max: 0 < 10081.5 < 381868.5 IQR (CV) : 20732.4 (1.7)	601 distinct values		743 (11.89%)	5506 (88.11%)
12	SLDPEA [numeric]	Mean (sd) : 12291.2 (22083.4) min < med < max: 0 < 3600.1 < 202050.6 IQR (CV) : 15676.2 (1.8)	335 distinct values		479 (7.67%)	5770 (92.33%)
13	DETBLO [numeric]	Min : 0 Mean : 0.2 Max : 1	0: 4851 (77.6%) 1: 1398 (22.4%)		6249 (100%)	0 (0%)
14	SLDBLO [numeric]	Mean (sd) : 14296.1 (29366.9) min < med < max: 0 < 2335.1 < 381868.5 IQR (CV) : 16065.2 (2.1)	819 distinct values		1398 (22.37%)	4851 (77.63%)

Figure 5: Descriptive statistics on the last group of continuous variables

2.3 Variables with NAs

From the above table descriptions, we realize that there are 4 variables containing missing values (NA). The table of these variables by proportion of missing values is given table 2.

Variables	Proportion of NAs
SLDLIQ	55.45%
SLDTIT	88.11%
SLDPEA	92.33%
SLDBLO	77.63%

Table 2: Table of variables with NAs

These variables correspond to balance of financial accounts (locked account, PEA, trading account, saving account) that some (or most) customers don't have. We thus imputed 0 in place of these NAs when categorizing.

2.4 Categorization of the continuous variables

In order to conduct our bivariate analysis in the following section, we categorized our continuous variables into 4 categories using their *min* and *max* values and their quantiles q_1 , q_2 and q_3 at

25%, 50% and 75% levels respectively. We created a 5th category = 0 for NAs.

- *category* = 0 if *var* = NA,
- *category* = 1 if *var* ∈ [*min*(*var*), *q*₁(*var*)[,
- *category* = 2 if *var* ∈ [*q*₁(*var*), *q*₂(*var*)[,
- *category* = 3 if *var* ∈ [*q*₂(*var*), *q*₃(*var*)[,
- *category* = 4 if *var* ∈ [*q*₃(*var*), *max*(*var*)[.

3 Bivariate Analysis

3.1 Pearson's chi square χ^2 measure of independance

The aim of this section is to make the bivariate analysis between *Y* and our explanatory variables. We would like to identify the explanatory variables most associated with *Y*. For this, we compute the Pearson's chi square χ^2 measure of independance for each variable.

The dependant variable has 2 categories. Let *k* be the number of categories of the explanatory variables.

$$\chi^2 = \sum_{i=0}^1 \sum_{j=1}^k \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Where $O_{i,j}$ is the Observed Frequency for which the explanatory variable = *j* and *Y*=*i*; and $E_{i,j}$ the theoretical expected Frequency for which the explanatory variable = *j* and *Y*=*i*, given the hypothesis of independance between the explanatory variable and *Y*. The results are displayed in Figure 6 for the binary variables and Figure 7 for the categorized continuous variables.

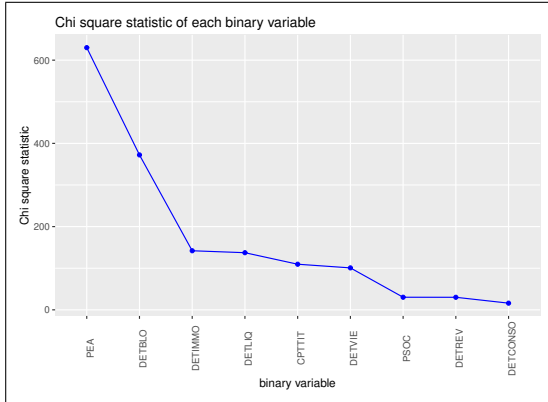


Figure 6: χ^2 measure between Y and all binary variables

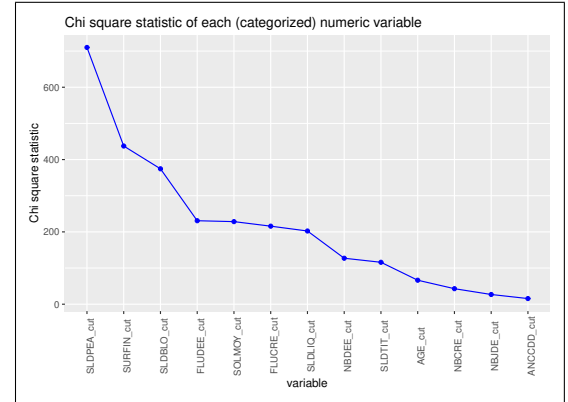


Figure 7: χ^2 measure between Y and all categorized continuous variables

Conclusion : The binary variables most associated with Y are PEA and DETBLO. The categorized continuous variables most associated with Y are SLDPEA, SURFIN, SLDBLO. We now study more in detail the relation of these variables with Y .

3.2 Proportion of Subscribers by Category

binary variables

PEA	subscribers (Y=1)	Proportion
0	742	13%
1	273	57%

Table 3: Proportion of subscribers by PEA

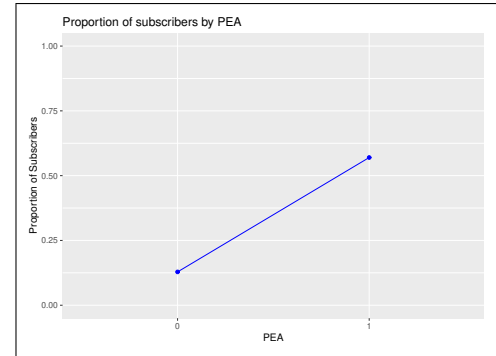


Figure 8: Proportion of subscribers by PEA

Owning a PEA account increases the probability of subscribing to the new Fund by 44 % points.

DETBLO	subscribers (Y=1)	Proportion
0	553	11%
1	462	33%

Table 4: Proportion of subscribers by DETBLO

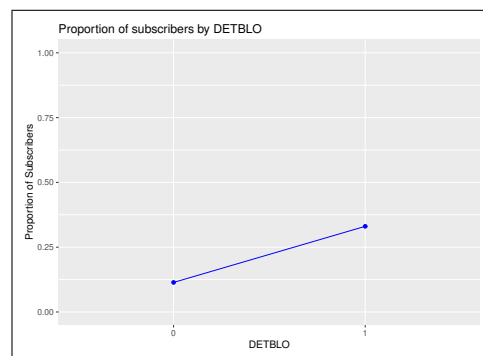


Figure 9: Proportion of subscribers by DETBLO

Owning a Locked saving account increases the probability of subscribing by 22 % points.

Categorized continuous Variables

SLDPEA	subscribers (Y = 1)	Proportion
0	742	13%
]0 , 10]	99	82.5%
]10 , 3.6×10^3]	59	49%
] 3.6×10^3 , 1.57×10^4]	56	47%
] 1.57×10^4 , 2.02×10^5]	59	49%

Table 5: Proportion of Subscribers per SLDPEA category

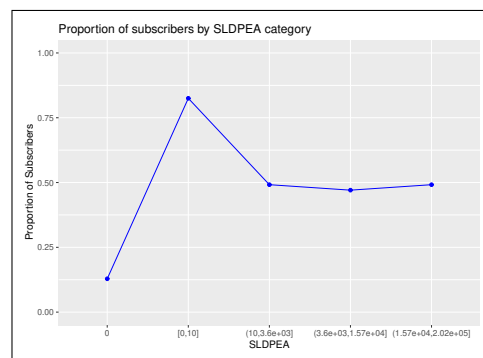


Figure 10: Proportion of Subscribers by SLDPEA category

The probability of subscribing increases by almost 70 % points between owning no PEA and owning a PEA with less than 10 euros, and stabilizes slightly below 50 % for higher amount.

SURFIN	subscribers (Y = 1)	Proportion
$[0, 26.4]$	65	4%
$]26.4, 2.63 \times 10^3]$	170	11%
$]2.63 \times 10^3, 1.46 \times 10^4]$	309	20%
$]1.46 \times 10^4, 6.54 \times 10^5]$	471	30%

Table 6: Proportion of Subscribers per SURFIN category

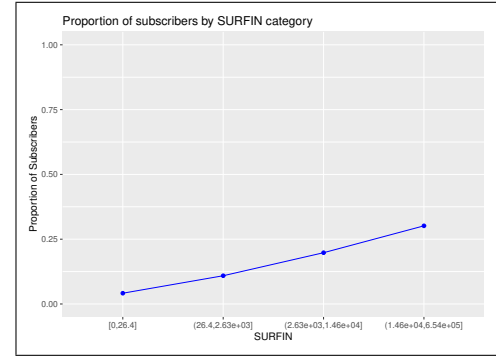


Figure 11: Proportion of Subscribers by SURFIN category

The probability of subscribing strictly increases with the amount of Total savings.

SLDBLO	subscribers (Y = 1)	Proportion
0	553	11%
$]0, 2.34 \times 10^3]$	232	33%
$]2.34 \times 10^3, 1.61 \times 10^4]$	111	32%
$]1.61 \times 10^4, 3.82 \times 10^5]$	119	34%

Table 7: Proportion of Subscribers per SLDBLO category

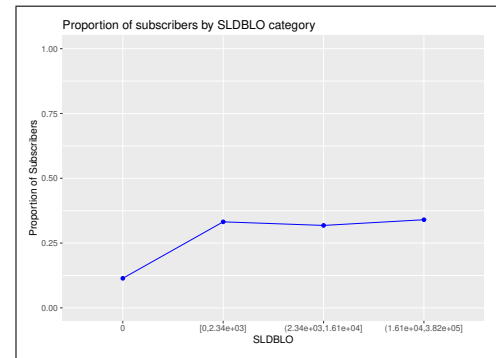


Figure 12: Proportion of Subscribers by SLDBLO category

Owning a Locked saving account increases the probability of subscribing by 22 % points, and this probability seems to remain stable regardless of the amount on the account.

The "Produits" variable

We create an additional variable named "Produits" which corresponds to the total number of financial products (accounts) the customer already owns in the bank, and which is strongly related to Y.

Produits	subscribers ($Y = 1$)	Proportion
0	113	6.25 %
1	224	11.3 %
2	284	22.4 %
3	197	26.2 %
4	129	41.3 %
5	46	54.1 %
6	18	51.4 %
7	3	100 %
8	1	100 %

Table 8: Proportion of subscribers by total number of products

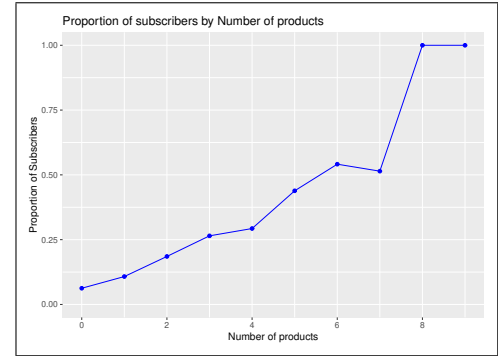


Figure 13: Proportion of subscribers by total number of products

The probability of subscribing roughly increases with the number of financial products (accounts) the customer already owns.

4 Prediction

The aim of this section is to predict the probability to purchase the new product Y . We use Random Forest and Support Vector Machine (Séparateur à vaste marge) algorithms, and compare the obtained results.

4.1 Random Forest

Random forest is an ensemble of decision trees algorithm based on the law of large numbers which can be used for classification. Each tree is fitted on a bootstrap sample and at each node the cutoff is selected among a subset of the explanatory variables drawn at random, to increase variability between trees. At the end of the algorithm, the predictions of all trees are averaged, and this predicted probability converges to the true probability according to the law of large numbers.

In this project, we fitted a random forest with 300 trees and let R default (which is reasonable) for the number of explanatory variables to draw at each node.

4.1.1 Performance on training data

Prediction\Reality	0	1
0	5234	23
1	0	992

Table 9: Confusion matrix on the training data

The confusion matrix on training data brings out a True positive rate or sensitivity equal to 97.7%, and a True negative rate or specificity equal to 100%. The ROC curve and lift (concentration) curve Figures 15 and 14 display these measures for varying threshold. The shape of the curves indicate that a perfect discrimination is possible for some threshold slightly lower than 0.5, for which the False positive rate remains to 0 and the True positive rate reaches 100 %.

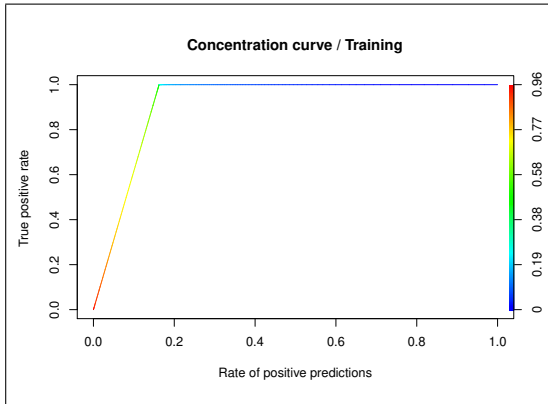


Figure 14: Concentration Curve / Training

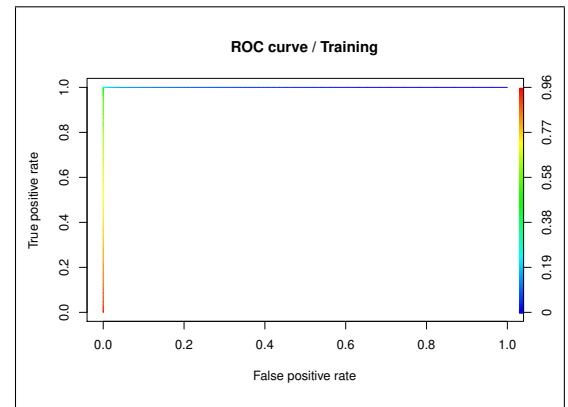


Figure 15: ROC Curve / Training

4.1.2 Performance on testing data

Prediction\Reality	0	1
0	1030	148
1	36	35

Table 10: Confusion matrix on the testing data

On testing data we note a significant decrease of the performance, with a True positive rate of 81% and a True negative rate of 97% on table 10.

Figure 16 and 17 display the concentration curve and the ROC curve. The concentration curve is often used in Marketing as it focuses on the True positive Rate, for a varying threshold. when the threshold is high, very few observations are predicted as positive, and these observations with highest scores (or probabilities) are likely to be in reality positive. However all true positives have not been selected if the threshold is placed too high, maintaining the True positive rate low. This true positive rate is then increasing with the Predicted positive rate, but theoretically slower and slower as the probability of the selected observations to be positive decreases with the threshold. But the more observations are selected, the more the true positive rate increases. When all positive observations have been selected, the curve reaches 100% and becomes flat. The "faster" the concentration curve reaches 100% and becomes flat, the better the model.

The concentration curve allows to determine the necessary proportion of the sample to select (to conduct the marketing campaign on) in order to achieve a desired True positive rate. In our case, if the bank wishes to reach 80% of the potential subscribers, she at least needs to conduct its marketing campaign on only 40% of the customers (the ones with the highest predicted probability obviously).

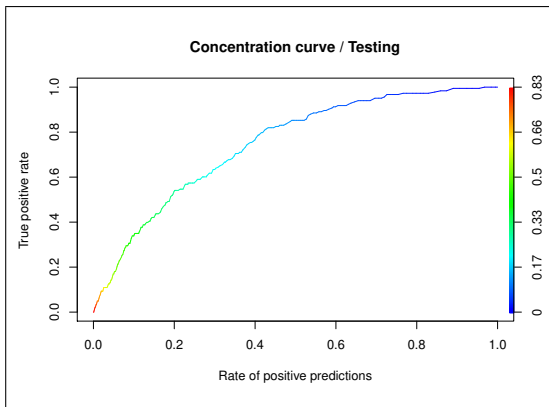


Figure 16: Concentration Curve / Testing

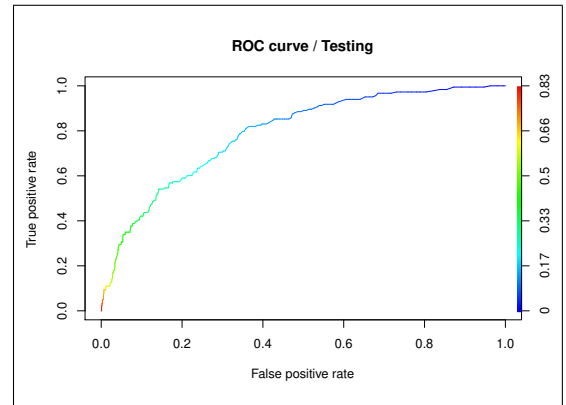


Figure 17: ROC Curve / Testing

4.2 Support Vector Machine

The Support Vector Machine algorithm is an optimisation algorithm based on the research of the optimal hyperplan in the space of explanatory variable in term of separation between classes of the dependant variable. This optimal hyperplan is the one which minimizes the number of observations wrongly classified; and for the rightly classified observations maximizes the margin between itself and the closest observations.

4.2.1 Performance on training data

Prediction\Realty	0	1
0	5116	754
1	118	261

Table 11: Confusion matrix on the training data

Table 11 shows that the algorithm achieves a 74% true positive rate and a 96% True negative rate on training data. The ROC and Concentration curves are displayed Figures 18 and 19.

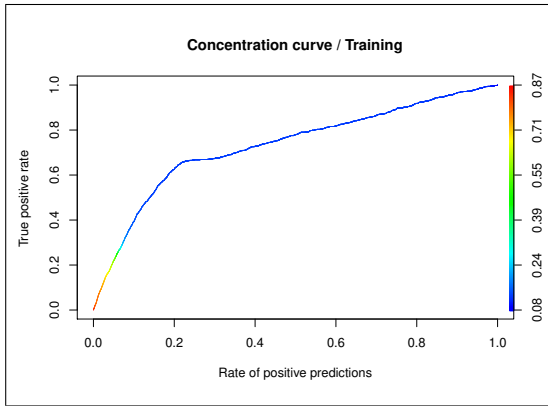


Figure 18: Concentration Curve / Training

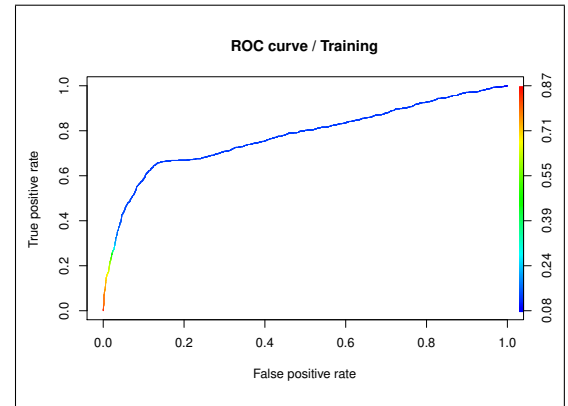


Figure 19: ROC Curve / Training

4.2.2 Performance on testing data

Prediction\Realty	0	1
0	1011	157
1	33	48

Table 12: Confusion matrix on the testing data

Table 12 brings out a 76% True positive rate and a 96 % True negative rate on testing data. These results are very close to the ones obtained on training data, unlike with the random forest algorithm. Random forest is a complex algorithm prone to overfitting the training data, whereas SVM has a degree of complexity lower, and doesn't overfit. By underfitting (or exactly fitting), the results on

testing data don't vary much from the ones on training data. The ROC and Concentration curves Figures 20 and 21 are very similar as well.

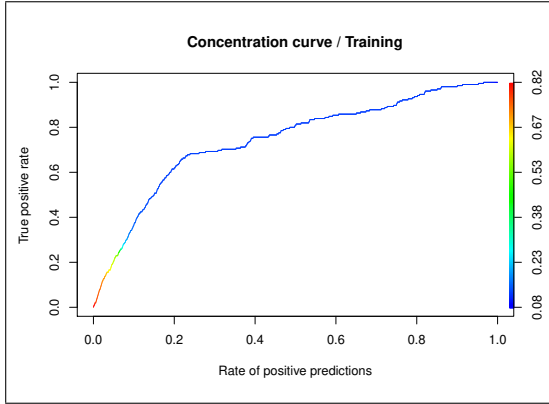


Figure 20: Concentration Curve / Training

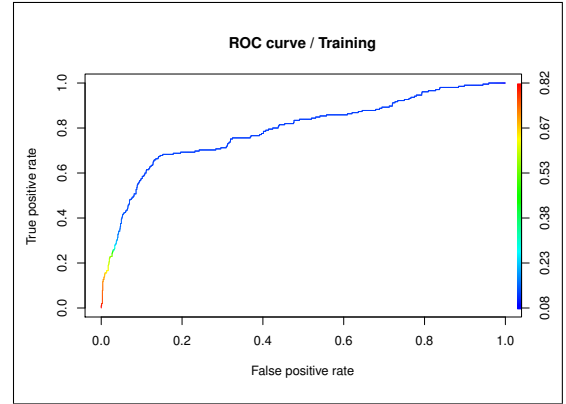


Figure 21: ROC Curve / Training

With SVM, 60% of the sample is needed to achieve a 80% True positive rate, which is higher than for the Random Forest (40%).

4.3 Algorithms Comparison

Figures 22 and 23 display the Rate of true responders within 10% wide categories of predicted probabilities. We note on these graphs as well that random forest performs better than SVM, with Frequency higher than SVM in the highest predicted probabilities categories, and lower than SVM in the lowest predicted probabilities categories. On the validation sample Figure 23, we note that it achieves a Rate of true responders beyond 80% in the highest category, while SVM only achieves 60 % in the same category.

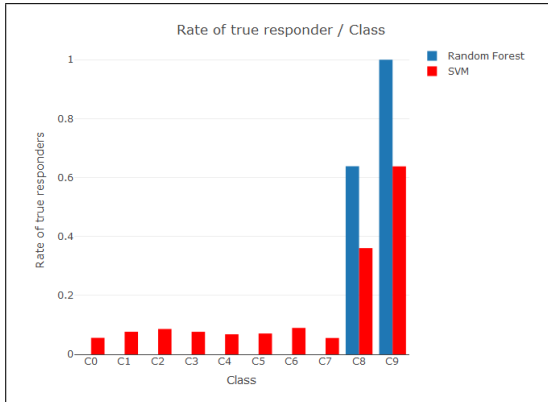


Figure 22: Rate of true responders / Class (Training sample)

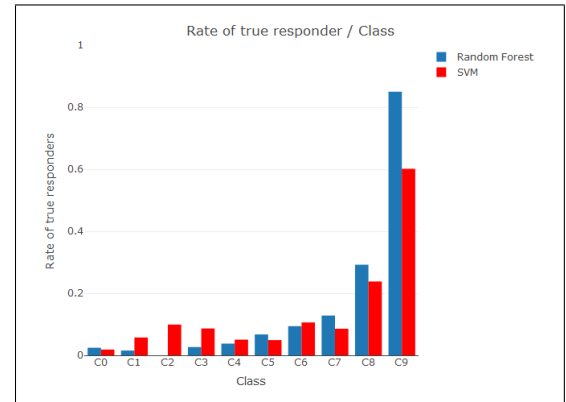


Figure 23: Rate of true responders / Class (validation sample)

5 Conclusion

The goal of this project was to predict the probability of each customer to purchase a new Financial Fund if proposed, sought. The end goal was to categorize customers in 10 groups according to their predicted probability, compute their average predicted probability (Return rate); and from this Return rate, taking into account Fixed and variable cost, fees, and Turnover, estimate the efficiency (net profit) of soliciting each group in a Marketing campaign. From this analysis the optimal customers for the bank to canvass can be deduced.

In our project, we identified the significant explanatory variables in predicting the customer probability to purchase, and assessed the performances of 2 algorithms in making this prediction. The significant results are:

- the probability of a customer to purchase increases with the number of (financial) products he already owns and with his amount of Total savings. Owning a PEA or a locked saving account also increases this probability.
- The random Forest algorithm performs better than the SVM algorithm. It achieves a 81 % True positive rate and a 97 % True negative rate on testing data. It allows to reach 80 % of the potential customers by canvassing only 40% of the sample.